Forecasting Violent Extremist Cyber Recruitment

Jacob R. Scanlon and Matthew S. Gerber

***Abstract*— The Internet's increasing use as a means of communication has led to the formation of cyber communities, which have become appealing to violent extremist (VE) groups. This paper presents research on forecasting the daily level of cyber-recruitment activity of VE groups. We used a previously developed support vector machine model to identify recruitment posts within a Western jihadist discussion forum. We analyzed the textual content of this data set with latent Dirichlet allocation (LDA), and we fed these analyses into a variety of time series models to forecast cyber-recruitment activity within the forum. Quantitative evaluations showed that employing LDA-based topics as predictors within time series models reduces forecast error compared with naive (random-walk), autoregressive integrated moving average, and exponential smoothing baselines. To the best of our knowledge, this is the first result reported on this forecasting task. This research could ultimately help assist with efficient allocation of intelligence analysts in response to predicted levels of cyber-recruitment activity.**

***Index Terms*— Violent extremist cyber-recruitment, forecasting, time series analysis, natural language processing.**

I. Introduction

THE modern landscape of extremism has expanded to encompass the Internet and online social media [1], [2]. In particular, violent extremist (VE) organizations have increasingly used these technologies to recruit new members. Within this article, a VE group is an organization that uses violent means to disrupt an established authority. Insurgents and terrorists are common types of VE groups that act with the specific goal of influencing public opinion or inciting political change. A radical religious group organizing inflammatory yet peaceful protests or a politically motivated person engaging in civil disobedience are not considered violent extremists under our definition. Many modern groups (e.g., the Westboro Baptist Church [3]) have radical religious views, but these beliefs are neither necessary nor sufficient to classify them as violent extremists without the intent to carry out or advocate for specific acts of violence. Within this article, VE recruitment is any attempt by a VE member to radicalize or persuade another person to aid his or her VE movement. VE cyber-recruitment is therefore VE recruitment activity that makes use of computers and the Internet.

Recent research by Torok shows that cyber tools are most influential at the onset of a future member's extremist activity—the recruitment and radicalization phase [2]. Extremist groups use the free and open nature of the Internet to form online communities [4] and disseminate literature and training materials without having to rely on traditional media outlets, which might censor or change their messages [2], [5]. Extremist organizations engage in directed communication and advertisement, recruiting members on social websites like Second Life, Facebook, and radicalized religious web forums [1], [2], [6]. Counter-VE organizations would benefit from an automatic method of forecasting recruitment activity within these online communities (e.g., number of recruitment posts per day in a discussion forum). Such a method would assist analysts in their search for such content and provide valuable information for efficiently allocating scarce human resources to counter the anticipated level of recruitment activity.

We hypothesize that authors of VE cyber-recruitment content do not behave randomly, but instead target forums and time periods in which the online communities appear to be vulnerable to recruitment propaganda. For example, a community might be vulnerable to VE recruitment if it is currently expressing dissatisfaction with the established authority. We further hypothesize that the semantic content (i.e., linguistic meaning) of messages is the key to expressing such vulnerabilities, as opposed to other linguistic levels such as morphology (word structure) and syntax (sentence structure). Our hypothesized links between community message content and recruitment activity lead to the following research question: How should the semantic content of an online community's digitized text be incorporated into forecasting models of VE cyber-recruitment?

We take up our research question in the following sections, which are organized as follows. In Section II, we compare off-line violent extremist recruitment with recent cyber-recruitment efforts. Additionally, we discuss previous counterinsurgency efforts and contemporary research that outlines the challenges associated with analyzing VE activities such as recruitment. In Section III, we describe the specific data sources used in our study, the text analysis and classification steps used to identify recruitment messages, and the methods used to compile the recruitment messages into a time series dataset. In Section IV, we define our time series models and explain how we incorporated the semantic content of digitized text into these models. In Section V, we compare the performance of these models to benchmark time series models and show that the former provide significant accuracy gains over the latter. Finally, in Section VI, we discuss potential improvements and future research

Manuscript received August 28, 2014; revised March 6, 2015 and May 30, 2015; accepted July 6, 2015. Date of publication August 5, 2015; date of current version September 15, 2015. This work was supported by the U.S. Army Research Laboratory under Grant W911NF-10-2-0051. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vrizlynn L. L. Thing. *(Corresponding author: Matthew S. Gerber.)*

J. R. Scanlon is with Teradata, Dayton, OH 45342 USA (e-mail: scanlon.jr@gmail.com).

M. S. Gerber is with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: msg8u@virginia.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2015.2464775

directions, including potential applications of our forecasting methods.

II. Related Work

A. Off-Line Recruitment

Many current jihadist insurgencies in Iraq and Afghanistan operate among local civilian populations and engage in both legal and illegal activities in order to achieve their strategic and political goals [7]. However, these activities are only effective when carried out by an organized and well-manned group [7]. Recruiting new members is thus a critical activity for both daily operations and the underlying political cause. An average terrorist group has a life expectancy of less than a year, so groups wishing to persist must replace members lost through arrests, deaths, and defections [4]. Several studies have investigated the causes of active participation within violent rebellions [8]–[12] as well as passive support for such movements [13]–[16]. The present article facilitates such understanding by investigating methods of extracting and using leading indicators of recruitment spikes and dips within online communities.

Ralph McGehee observed VE recruitment during his 1967 work to identify communist insurgents in the rural villages along the northern border of Thailand. His efforts helped the joint CIA-Thai counterinsurgency to more accurately forecast aid demand within at-risk villages, thereby improving regional support for the Thai government. The success of McGehee's program is attributed to his intelligence teams collecting information on all individuals in the villages, not just on the insurgent sympathizers. This provided a more complete picture of the community and permitted more accurate forecasts of insurgency support [17]. Although this article investigates forecasting within online communities, strong parallels exist between these virtual worlds and the physical communities addressed by McGehee; both contain VE groups that operate within, hide among, and recruit from a population of potentially sympathetic individuals.

B. Cyber-Recruitment

The primary strength of cyber-recruitment is its ability to quickly expose large online communities to a substantial amount of engaging multimedia content [2], [18], [19]. Much of the research on cyber-recruitment within these communities has focused on how violent groups use legitimate social networking websites along with online discussion forums for recruitment and other activities. This prior research largely provides evidence and case studies of online VE activity and suggests ways that virtual worlds may be used by these groups in the future [5], [6], [19]–[22]. Recent research has evaluated the use of political tools for shutting down websites or shaming material supporters [23]. Some researchers have suggested the use of web-crawling and analysis techniques to monitor for VE activities including recruitment [2], [24], [25]; however, there do not appear to be any implementations of such techniques. The present article describes research that fills this gap, addressing the need to forecast cyber-recruitment activity in online social media settings.

Computer-based social network analysis (SNA) is a large field of research, one objective of which is to identify the organizational structure of VE networks. With objectives similar to McGehee's manual SNA work, contemporary research aims to detect the presence of VE groups and their influence within large-scale networks based on the number of interconnections among VEs and influential community members [1], [26]–[29]. There have also been preliminary attempts to profile individual users using text mining techniques [30]. However, this prior research has typically focused on violent extremist activity in general without focusing on recruitment.

In sum, although previous research has investigated aspects of VE cyber-recruitment as well as computational approaches to network analysis, textual analysis has only been investigated in a preliminary way and such techniques have not been connected with time series forecasting methods to forecast VE cyber-recruitment activities in online communities. This article complements the research surveyed above by building on recent data collection efforts, focusing on online recruitment specifically, and applying current techniques from natural language processing and time series analysis to forecast recruitment activities.

C. Time Series Analysis of Digital Text

Researchers have investigated a variety of ways to analyze the temporal evolution of digitized document collections. The present article focuses on topic modeling, which is an unsupervised, generative probabilistic model of document collections that reveals broad topical themes within documents [31]. We will explain topic modeling in more detail in Section IV-A. For now, suffice it to say that a topic is a cluster of words that captures one theme within a document collection. For example, within online discussions of violent extremism, one would expect to see separate clusters of words describing violent acts and rewards for participation. The topic modeling process identifies these clusters of words following an approach that is conceptually similar to unsupervised clustering.

A key assumption of early topic modeling algorithms is that each document in the collection is exchangeable [31]. That is, the order of documents within the probabilistic model has no effect on the objective function being optimized. For collections of documents that either have no timestamp or have no temporal evolution, this is a reasonable assumption; however, certain collections and certain analyses require consideration of temporal evolution. For example, when studying the evolution of a particular scientific field, it is crucial to place published articles on a timeline that captures the relative ordering of ideas. Such articles are not exchangeable if one's goal is to understand the evolution of topics within the field. In response to this need, Wang and McCallum proposed the Topics over Time (TOT) model, which captures the waxing and waning of topic prominence within a document collection over a continuous timeline [32]. Blei and Lafferty discretize time, analyze topics within each time interval, and require topics in one time interval to evolve from topics in the previous interval [33]. Each of

TABLE I
FORUMS USED IN OUR STUDY, EXTRACTED FROM THE ANSAR ALJIHAD NETWORK VIA THE DARK WEB PORTAL

	AsAnsar	Ansar1
Time-frame	11/2008 - 5/2012	12/2008 - 1/2010
Messages	269,548	29,492
Members	5,034	382
Language	Arabic	English

these approaches makes structural changes to Blei et al.'s original generative model [31]. In contrast, Hall et al. run a simple topic analysis on each time interval and perform *post hoc* analysis on the topic strengths to glean insight about the progression of topics within the documents.

The present article builds on the work of Hall et al. in that we do not modify the generative probabilistic model formulated by Blei et al. [31]. We depart from this work, however, in that we conduct *post hoc* analysis with formal time series methods such as ARIMA. Our experiments demonstrate the advantage of our *post hoc* time series analysis over baseline ARIMA and exponential smoothing models that simply model the time series of interest (frequency of VE recruitment posts), ignoring contextual information provided by the content of such posts (i.e., their topics).

III. DATA COLLECTION AND ANNOTATION

Given our interest in VE recruitment, we focused our investigation on so-called "dark web" content. Dark web content is defined as information from typically private social websites where extremists interact. Many early efforts focused on locating, accessing, extracting, and storing data from dark web forums [1], [24], [34]. In previous work we leveraged these efforts to build supervised classifiers that automatically identify VE recruitment posts [35]. Below, we briefly review our dark web data sources and our previous investigation into automatically identifying recruitment messages. The present article uses the resulting identified posts to build forecasting models for VE recruitment activity within dark web forums.

A. Data Sources

In previous research [35], we identified the Dark Web Portal Project [36] as an ideal data source for research on violent extremist recruitment. The Dark Web Portal is a repository of social media messages compiled from 28 different online discussion forums. These forums focus on extremist religious (e.g., jihadist) and general Islamic discussions, many of which are sympathetic to VE movements. Most of the thirteen million collected messages come from Arabic sources, but the Dark Web Project provides translation services and compiles information from at least seven dedicated English-language forums. The most relevant forums come from the Ansar AlJihad Network, which we summarize in Table I and describe in more detail below.

The Ansar AlJihad Network is a set of invitation-only jihadist forums in Arabic and English that are known to be popular with western jihadists [37]. The Dark Web Project compiled 299,040 total messages posted on Ansar AlJihad between 2008-2012. Fewer posts are compiled from the English forums, called Ansar1, than from the Arabic portion of the site; however, the English subset was sufficiently large for our study and contained contemporary, original-English discussions between jihadists and jihadist sympathizers. We used this subset in all of our experiments.

B. Data Pre-Processing and Classification

We collected and pre-processed the Ansar1 posts into time series data as follows. We first ingested all 29,492 raw Ansar1 forum posts and compiled the message text and respective message IDs into an initial corpus. We then automatically removed duplicates (same message ID) and empty documents (no message text), retaining 28,744 posts in the corpus. Most posts contained exclusively English text, as Asnar1 is the English-language forum for the Ansar AlJihad Network. However, occasional posts included non-English words or phrases; these were typically Arabic passages from the Koran. In these cases the non-English passages were converted to English using Google Translate [38]. We left slang words written in Latin characters intact under the assumption that they were meant to be readable by an English language speaker. For example, "Kuffar" is a derogatory Arabic term for non-believer.

The Dark Web Portal project does not indicate which messages contain VE recruitment content and which do not. Thus, we manually annotated each post within a sub-corpus of 294 randomly selected posts, adding a binary variable indicating recruitment or non-recruitment. Examples of annotated posts are shown in Table II. Agreement analysis with Cohen's κ between two annotators indicated $\kappa = 0.70$ with confidence interval $(0.5, 0.7)$ at $p = 0.01$ (significant non-random agreement). We then used these annotated posts to build and evaluate supervised naive Bayes, support vector machine (SVM), and boosting classifiers. Our SVM classifier demonstrated an area under the receiving operating characteristic curve (AUC) of 89% on the task of recruitment post classification. Further details of our approach are provided in [35].

In the present research, we used our best SVM classifier to automatically annotate the entire Ansar1 corpus, producing a final set of 28,744 posts each with a label of recruitment or non-recruitment. We sorted these posts in ascending order by date and compiled a count of total posts and the number of VE recruitment posts each day. These daily counts were used to measure the raw number of recruitment posts per day as well as the percentage of all posts per day that were recruitment (see Figure 1). Due to the large data gaps seen in early segments of the timeline in Figure 1, we created a continuous time series by removing all posts from dates prior to January 14, 2009. The exact cause of these gaps is unknown, but they could be due to several factors including a temporary forum shutdown, or the data-collection mechanism losing access to the website.[1]

[1] Personal communication with Dark Web technicians.

TABLE II

EXAMPLE TEXT OF ANSAR1 FORUM POSTS AND THE RESPECTIVE ANNOTATIONS

Annotation	Sample Text
Recruitment	*A Golden chance to join Jihad in Somalia. Abo Dojana invited those who want to participate in jihad to join the militants in Somalia to form what he called a base of martyrdom-seekers who would from there spread to the entire world. Somalia could actually be an ideal base for physical and weapons training...*
Recruitment	*Representing the militant Islamic group Shebab, Abu Mansour makes a pitch for new overseas recruits after praising one militant fighter killed in an apparent ambush. 'So, if you can encourage more of your children and more of your neighbors and anyone around to send people like him to this jihad (holy war), it would be a great asset for us,' he says.*
Not Recruitment	*I have now added him as a friend on Facebook. But something tells me that he isn't going to answer to my request. LOL, you had me rolling on the floor man!!!!! So this attack was done my 'Jaish al Mujihadeen' How it that possible? , did they have problems with bounced-checks from the US?*
Not Recruitment	*A court in the German city of Koblenz sentenced a German of Pakistani origin to eight years in prison Monday on a conviction of assisting the international Al-Qaeda terror network. The man gave the group financial aid and tried to recruit new members in German territory, according to the indictment*
Not Recruitment	*Did Mansoor join the emerat? I heard he is still fighting for Ichkira Republic*

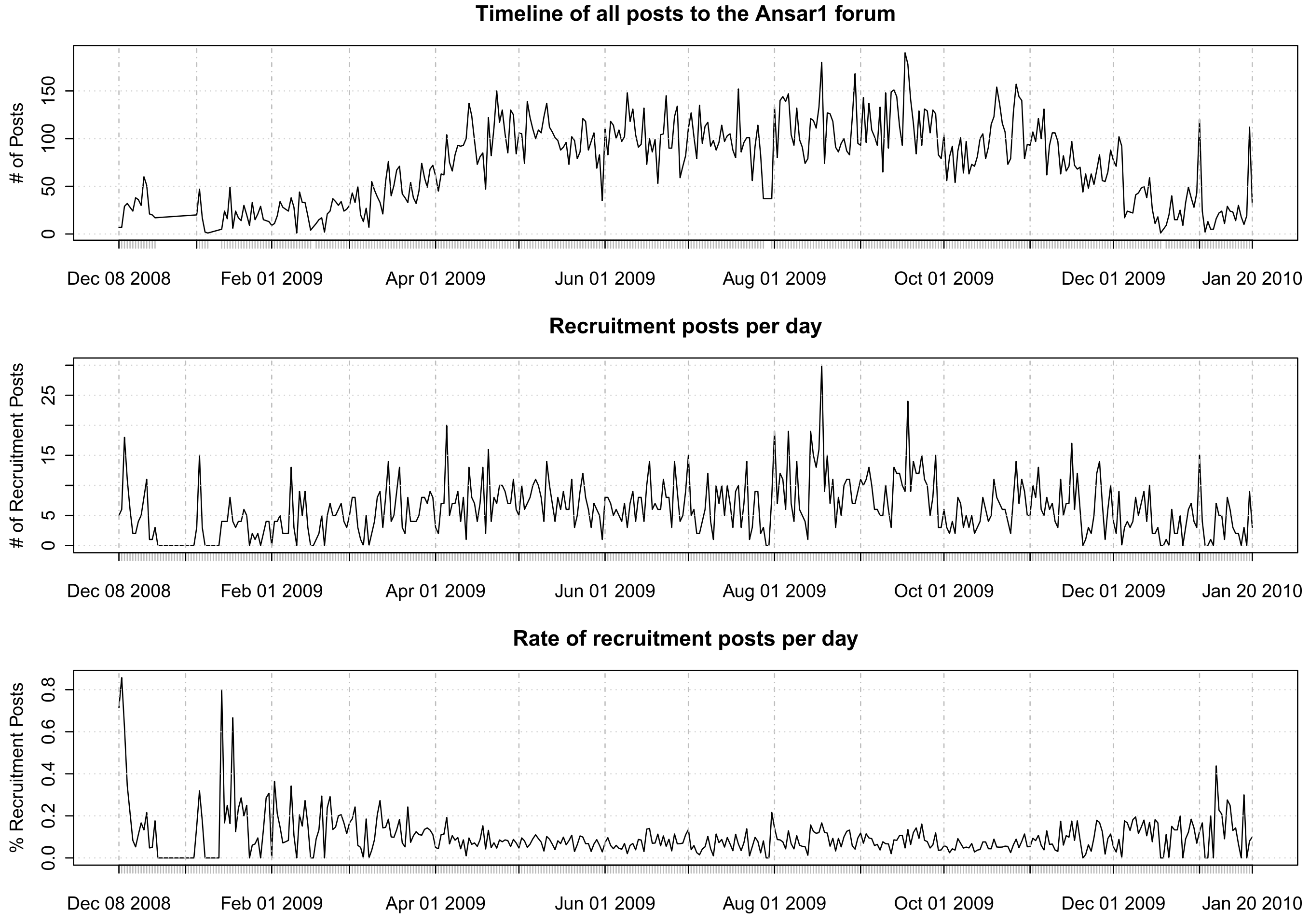

Fig. 1. Timeline of forum posts compared to the response variables: (top) Ansar1 forum activity, (middle) the daily count of VE recruitment posts, and (bottom) the daily percentage of VE recruitment posts among the total volume of posts.

We modeled two response variables describing the level of VE recruitment activity in our corpus: (1) the number of VE recruitment posts per day, and (2) the percentage of posts containing VE recruitment per day. Figure 1 shows a timeline of these response variables compared to the total level of activity on the Ansar1 forum during the 2009 calendar year. In the following sections, we present different forecasting methods for these two time series response variables.

IV. ANALYTIC APPROACH

We developed several methods to forecast the count and percentage of VE recruitment posts. We used the following general model:

$$E[Recruitment_t \mid d_{t-1}] = F[x_1(d_{t-1}), \ldots, x_n(d_{t-1})] \qquad (1)$$

In Equation 1, $Recruitment_t$ is a continuous response representing the level of VE recruitment (either raw or percentage) expected at the current time period, d_{t-1} represents

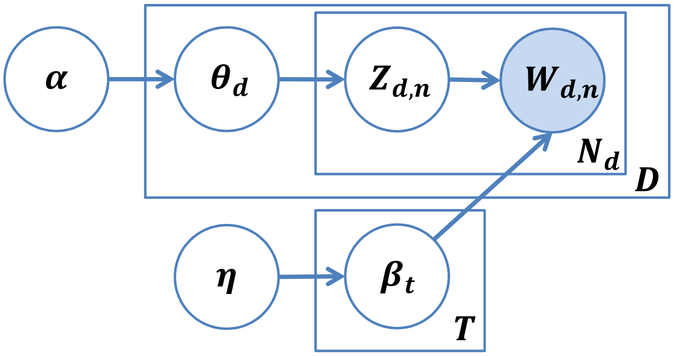

Fig. 2. Generative LDA topic model for Ansar1 posts.

the textual concatenation of all forum posts observed within the previous day, and x_j represents a feature function of d_{t-1}. In the following sections, we discuss the feature functions used in forecasting and present different formulations of the forecasting function F.

A. Topic Modeling for Ansar1 Forum Posts

We hypothesized that the textual content of posts from the prior day d_{t-1} would contain information that may aid forecasts of recruitment activity on day d_t. For example, forum members might discuss recent attacks by an established government just prior to initiating or increasing recruitment efforts online. To operationalize this intuition, we transformed the textual content associated with each day d_t into a numeric vector for use with time series forecasting methods. Specifically, we started with a bag-of-words transformation in which the text from all days is contained in a term-by-document matrix. In this matrix, each row is a term, each column is a document representing the forum content from a particular day, and each cell is the count of the given term on the given day. Each document (column) is thus a high-dimensional vector of term frequencies, and the task is to transform the high-dimensional matrix of document vectors into a low-rank approximation that captures key differences between documents (days). Note that this aggregation and transformation process does not change the response (dependent variable). It only affects the independent variables, which are derived from the term-by-document matrix.

We created our low-rank approximation with topic modeling. Specifically, we used latent Dirichlet allocation (LDA), which reduces the high-dimensional term-by-document matrix into a topic-by-document matrix where rows are thematic topics, columns are documents (days), and cells indicate the strength of the given topic in the given document [31]. The standard LDA model is represented graphically in Figure 2. It is a hierarchical Bayesian model that extracts latent topic variables from a corpus of documents. LDA's generative process for each document d in a corpus D can be described as follows:

1) Draw T topic-word distributions, each a multinomial $\beta_t \sim Dir_V(\eta)$ where $Dir_V(\eta)$ is a symmetric, V-dimensional Dirichlet distribution, V is the number of rows in the original term-by-document matrix (i.e., the vocabulary size), and η is the symmetric prior on the Dirichlet distribution.
2) For each document d, draw a K-dimensional multinomial $\theta_d \sim Dir_K(\alpha)$, where $Dir_K(\alpha)$ is another symmetric Dirichlet distribution with prior α, and K is the number of topics to be modeled (selected by the user).
3) For each word $w_{d,n}$ in document d, draw a topic $z_{d,n} \sim Multinomial(\theta_d)$ and then draw a word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$.

After optimizing the topic-word and document-topic distributions, the latter describe the topical composition for each document in terms of topic proportions. This is the low-rank approximation of the original term-by-document matrix, and we used the values in this matrix as predictors $x_j(d_{t-1})$ in our forecasting models with j indexing a row in the matrix and $t-1$ indexing a column (i.e., the content of the forum on day $t-1$). This approach is effective in the setting of crime prediction based on social media [39], and we hypothesized its effectiveness on our task as well.

We implemented the above approach in the following way. First, we used the *RTextTools* and *tm* text mining packages in R to perform word stemming and remove URL web addresses, numbers, punctuation, stop words, and whitespace [40]–[42]. Since the forecasting approach uses discrete time periods, we grouped the posts by day, concatenated the message text of all posts in a day to form a daily "document", and generated $K = 30$ latent topics from these documents using the LDA algorithm implemented within the R package *topicmodels*.[2] We used the default prior topic weight of $\alpha_k = 1.67$ for all topics. We then used the topic proportions output for each day as predictors within our various time series models, which are described in the following section.

B. Time Series Models and Regression Functions

We designed our experimental setup to reflect the practical scenario in which a new prediction is made each day using all prior days' data (e.g., in order to drive daily staffing allocations of intelligence analysts). For example, an organization might counter an anticipated spike in VE recruitment by staffing additional analysts to investigate leads. Specifically, when predicting the response for day d_t, we trained a forecasting model on response values from all previous days. For example, when predicting the response on d_8, we used training responses from days $d_1 - d_7$. The predictor variables associated with one of these training responses were extracted from the days prior. So, for training response d_2, predictor variables were extracted from day d_1; for training response d_3, predictor variables were extracted from days d_1 and d_2; and so on. All training responses and their associated predictors were used to build

[2] We chose to concatenate forum posts into a daily "document" rather than analyzing each forum post separately, since the latter approach would require aggregation of each post's 30 topic probabilities into a daily vector of 30 topic probabilities. One might achieve the latter by averaging the topic probabilities of the posts, perhaps weighting by the post length; however, this is substantially more demanding in terms of LDA computation, and we suspect the end result would be quite similar to the simpler approach of aggregating first and then analyzing.

a supervised model to predict the response for day d_8. We used this prediction for d_8 and the associated ground truth to calculate forecasting error on day d_8. The prediction day moved sequentially across the timeline, repeating the above process at each step, to calculate aggregate performance scores for our models. This setup reflects the practical scenario in which an organization retrains its predictive models each day, incorporating one new day of information at each step.

Prior work on this particular forecasting task does not exist. Thus, we implemented three baseline approaches within the experimental setup described above: naive (random walk), autoregressive integrated moving average (ARIMA), and exponential smoothing (ETS). These baseline approaches operate solely on the basis of past VE recruitment response levels, ignoring any textual content and associated topic analyses. We then compared the performance of these baselines to approaches that include topic analyses, in order to assess the contribution of topic content to the forecasting task. Some of our topic-based approaches use traditional regression analysis; however, we also present methods that use topic content within time series models like ARIMA. Below we first present the baselines and then present the details of each topic-based approach.

Baseline 1 (Naive Model): Our naive model, also called a random walk, is a trivial forecasting method in which the forecast for each day d_t is the observed value for day d_{t-1}. It is called a naive model because it assumes that there is no change from one period to the next. We use this naive model as a basic benchmark for comparison against more advanced forecasting techniques.

Baseline 2 (Autoregressive Integrated Moving Average (ARIMA)): We also implemented an ARIMA model to forecast VE cyber-recruitment activity. Prior to fitting the model, integration (differencing) was applied if the time series data were not stationary (i.e., centered at zero). Given the stationary prior recruitment response values $Rec_1, \ldots, Rec_{t-1}$, we applied the following ARMA model with p autoregressive (AR) terms and q moving-average (MA) terms [43]:

$$E[Rec_t \mid \mathbf{Rec}_{1,\ldots,t-1}] = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i Rec_{t-i} + \sum_{i=1}^{q} m_i \varepsilon_{t-i} \tag{2}$$

In Equation 2, c is a constant intercept term, $\varepsilon_t, \varepsilon_{t-1}, \ldots, 1$ are residual error terms after removing trend and seasonality, and α_i and m_i are the parameters for the autoregressive and moving-average portions of the model, respectively.

We used the *forecast* package in R to automatically select the optimal number of AR and MA terms and then fit the model using OLS regression [44]. Only one order of differencing was required to make the time series stationary. All results labeled *Baseline ARIMA* in Section V were produced with an ARIMA model fit in this way. This basic ARIMA time series model is another benchmark against which more advanced models are compared to determine whether additional predictors, like latent topics from textual content, improve forecasting accuracy.

Baseline 3 (Exponential Smoothing (ETS)): Lastly, we implemented an ETS model that, like ARIMA, uses previous VE recruitment response values as predictors of future recruitment activity. Simple exponential smoothing consists of a weighted average of past observations with recent observations weighted higher than older observations. Our implementation of ETS used the state-space framework and automatic forecasting methods proposed by Hyndman et al. [45] and implemented in the *forecast* package [44]. All results labeled *Baseline ETS* in Section V were produced with an ETS model fit in this way.

Topic-Based Approach 1 (Principle Components Regression (PCR)): Figure 3 shows how we used PCA to obtain the predicted response for day d_t. Starting on the left, we first extracted 30 topic proportions from each day's document up through d_{t-1}. We then computed the principal components of this topic space to eliminate multicollinearity. To obtain a predicted response for d_t, we regressed the responses from days $d_2 \ldots d_{t-1}$ on the PCA topic proportions from days $d_1 \ldots d_{t-2}$, respectively, using standard linear regression (OLS). We then used the trained OLS model and the PCA topic proportions from d_{t-1} to predict the response for d_t. We determined the optimal number of PCA components via cross-validation on the training data, further eliminating model dimensions while retaining the highest variance PCs for our feature space. The PCR results shown in Section V as TopicPCR were produced using the R package *pls*, which generated the PCs of the 30 latent topic dimensions and selected the optimal number of PCs at each iteration of the moving prediction window [46]. This entire process is repeated for all days d_t to obtain overall performance scores.

Topic-Based Approach 2 (Topic Time Series): The PCR approach explained above implements the intuition that yesterday's topics of discussion provide indicators about today's level of VE recruitment. We also hypothesized that non-lagged topic probabilities might provide predictive power in our forecasting model. In other words, regressing the d_t response on topics from d_t might give better predictions. The trouble with this is that d_t refers to a future date, so the associated forum content has not yet been generated. Thus, we turned to time series methods to forecast the topic distributions on day d_t. We implemented two methods: autoregressive integrative moving average (ARIMA) and exponential smoothing (ETS). We call these two approaches "topic time series" models, an example of which is shown in Figure 4. The following steps were taken at each iteration of the moving prediction window:

1) We fit a PCR model using the 30 latent topics as described previously for TopicPCR. The only difference is that we regressed the responses from $d_1 \ldots d_{t-1}$ on the topic proportions observed on $d_1 \ldots d_{t-1}$, respectively. Since this model depends on the topic probabilities for the prediction day d_t, and since these probabilities cannot be known ahead of time, we forecasted the topic probabilities for day d_t as described in the next step.
2) We used ARIMA and ETS models to forecast the 30 topic probabilities on day d_t using the topic probabilities from days $d_1 \ldots d_{t-1}$. The forecasted

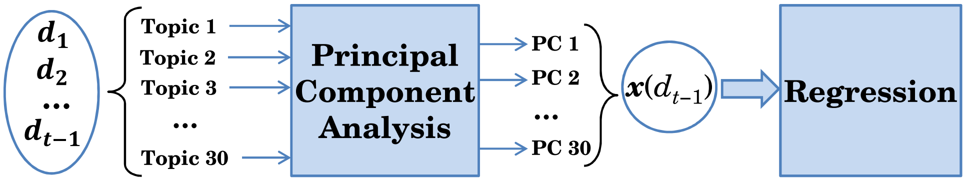

Fig. 3. PCR model layered on top of the LDA topic model. Input documents on the left create topic vectors via LDA, from which principal components are extracted and used in OLS regression.

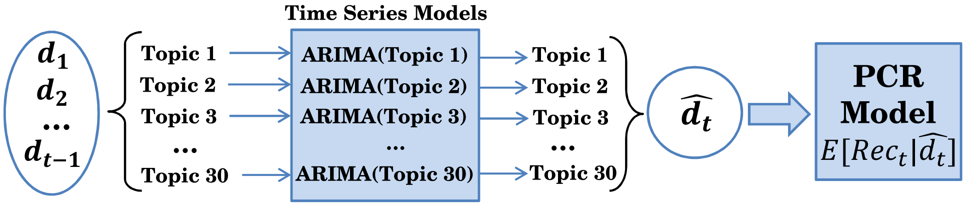

Fig. 4. ARIMA model layered on top of the LDA topic model and feeding into the PCR model from Figure 3. Input documents on the left create topic time series vectors via LDA (30 topic probabilities per day). Each of these 30 topic time series are modeled via ARIMA to forecast the topic probability on the prediction day d_t. The forecasted topic time series are reprojected into a principal component space via PCR, which is used to forecast VE recruitment activity via OLS regression.

topic probabilities for day d_t are shown in Figure 4 as $\widehat{d_t}$. We used one time series model per topic. This is depicted in the figure as $ARIMA(Topic1)\ldots ARIMA(Topic30)$.

3) We used the estimated topic probabilities $\widehat{d_t}$ from step 2 as inputs to the PCR model trained in step 1. This results in a forecast of the current day's VE recruitment activity using an estimate of the current day's topic probabilities ($E\left[Rec_t|\widehat{d_t}\right]$).

The results labeled *TopicARIMA* and *TopicETS* in Section V below were produced using this topic time series modeling method.

V. Results and Discussion

We rebuilt each of the above models on each day to obtain response predictions for all days in our data. We then tested for nonlinear serial dependence in our time series by running a BDS test on our prediction residuals. The results of this test for each model are described below:

- Baseline ARIMA: No evidence of nonlinearity.
- Baseline ETS: No evidence of nonlinearity.
- TopicPCR: Significant evidence of nonlinearity.
- TopicARIMA: Very little evidence of nonlinearity.
- TopicETS: Very little evidence of nonlinearity.

Given the above results, we continued with evaluation of all models. We discuss the nonlinearity of the TopicPCR model in Section VI.

We evaluated the above VE recruitment forecasting methods using two measures of forecasting accuracy: root mean squared error (RMSE) and mean absolute scaled error (MASE); these are depicted in Equations 3 and 4 below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(F_t - A_t)^2} \tag{3}$$

$$MASE = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{|F_t - A_t|}{\frac{1}{n-1}\sum_{i=2}^{n}|A_i - A_{i-1}|}\right) \tag{4}$$

TABLE III

MASE and RMSE Results for Time Series Forecasting of Total VE Recruitment Posts Per Day (# Per Day) and Percentage of Recruitment Posts Per Day (% Per Day)

	MASE		RMSE	
	# per day	% per day	# per day	% per day
Baseline 1: Naive Model	1.08	1.65	5.23	0.10
Baseline 2: ARIMA	1.05	11.37	5.29	0.70
Baseline 3: ETS	1.17	7.24	5.75	0.46
OLS	5.41	9.72	150.40	6.50
TopicPCR	0.87	0.90	4.55	0.06
TopicARIMA	0.84	1.16	4.39	0.07
TopicETS	0.86	0.90	4.48	0.06

In Equations 3 and 4, A_t and F_t are the actual and forecasted recruitment activity levels at time t, and the denominator of Equation 4 is the average absolute forecast error of the naive model. We used MASE because it can compare forecasting methods while not returning extremely high or even infinite error rates when forecasts are near zero [47]. RMSE is also helpful for understanding forecasting accuracy because the error rate is in the same units as the response.

Table III shows a performance comparison of all forecasting methods. These results were obtained by applying the RMSE and MASE metrics to the set of forecasts predicted at each iteration of the moving prediction window. First, we observe that neither the baseline ARIMA nor the baseline ETS models outperformed the naive baseline in terms of RMSE. With the addition of latent topics, however, these models are improved. TopicPCR, TopicARIMA, and TopicETS all improve forecast accuracy over the baseline models for each of the metrics. Examining RMSE shows that models of total recruitment posts per day and percentage of recruitment posts per day improve by 16% (Baseline 1 versus TopicARIMA) and 40% (Baseline 1 versus TopicPCR or TopicETS), respectively. These improvements are due to the incorporation of latent, text-based indicators extracted via LDA topic modeling. The full OLS models are shown in Table III for comparison with PCR. These OLS models use the topic proportions directly, skipping the PCR reprojection step. The substantial improvements

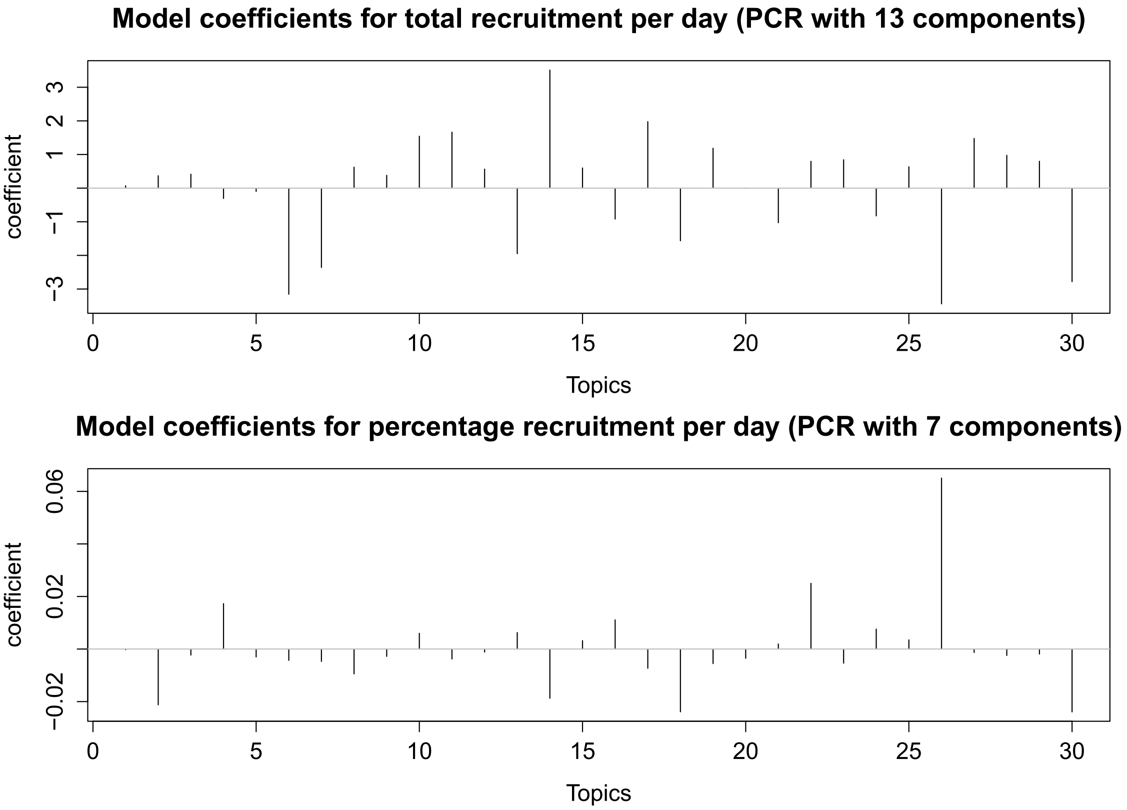

Fig. 5. Comparison of topic feature coefficients within the TopicPCR model. Total recruitment posts per day (top) and percentage of recruitment posts per day (bottom).

TABLE IV

TERMS CONTRIBUTING TO THE LOWEST-WEIGHTED (TOPIC 3) AND HIGHEST-WEIGHTED (ALL OTHER) TOPICS IN THE TOPICPCR MODEL

Topic 3	Topic 6	Topic 7	Topic 14	Topic 26	Topic 30
und	said	said	kill	allah	said
die	allah	allah	attack	will	kill
les	attack	islam	islam	one	attack
der	kill	will	said	muslim	police
des	quot(ed,ing,...)	god	soldier	people	milit(ant,ary,...)
que	brother	govern	taliban	jihad	taliban
qui	mujahideen	muslim	offic(ial,er,...)	brother	forc(e,ed,...)

of TopicPCR versus OLS demonstrate that multicollinearity and a lack of feature selection can dramatically increase prediction error. Lastly, regarding the absolute size of the improvements shown in III, we note that in a setting such as VE cyber-recruitment, conducting a follow-up investigation on a single recruitment post could be an expensive undertaking for the investigating analyst. Thus, reducing the error by even one post might be a practically significant achievement.

To provide some understanding of how the best-performing regression models are being trained to predict the level of VE recruitment activity, Figure 5 shows the importance of the latent topic features within the TopicPCR models for the two response variables. The models in Figure 5 represent the last moving prediction window, which uses the largest set of training data. The coefficient (y-value) for each topic (x-value) is computed as the sum of PCR loadings for the topic across the selected principal components, weighted by the coefficients assigned to the principal components within the regression. The coefficients show importance for topics 14, 18, 26, and 30 within each model.

Table IV provides some understanding of the most important topics for the dataset, including 14, 26, and 30, which were important within Figure 5. Comparing topic 26 to topics 14 and 30, the former appears to focus more on the religious aspects of conflict, whereas the latter appear to focus more on the physical aspects of conflict. Topics 6 and 7 feature prominently in the top sub-figure of Figure 5. We see a similar division between these two topics, with the former focusing on physical aspects of conflict and the latter focusing on religious aspects of conflict. These observations support our hypothesis that conflict and its attendant discussion precede the creation of VE recruitment content within online communities. Interestingly, we did not observe topics that focus on recruitment activities. This is likely due to the scarcity of VE recruitment posts, which comprise less than 10% of the Ansar1 corpus. Regardless, our performance metrics show that latent topic features improve VE recruitment forecast accuracy over the benchmark time series models.

Another key observation is that the forum corpus contains many posts that are not related to religion and conflict. For example, Topic 3 in Table IV contains non-content words from French, German, and Spanish. As shown in Figure 5, this intuitively irrelevant topic receives a small weight within the regression model and is effectively filtered out. Our empirical results are consistent with the hypothesis that religion and conflict are the drivers of the recruitment process.

VI. Conclusions and Future Work

This research was motivated by the continuing increase in online activities of violent extremist organizations along with the lack of automated tools to predict such activity. We have built upon recent data collection and analysis efforts to develop methods that automatically forecast VE cyber-recruitment using natural language processing, supervised machine learning, and time series analysis. Our results indicate that automatic forecasting of VE cyber-recruitment is a feasible goal. As the first reported results on this task, our time series models serve as initial performance benchmarks against which future models can be compared. The data used in our experiments are available online.[3]

The methods we have developed could eventually be integrated into analyst workflows to help alleviate the burden imposed by "ever-increasing [volumes of] information" stored in intelligence databases [48]. This could be accomplished in two steps. First, the forecasting methods presented in this article could be used to predict future levels of VE cyber-recruitment activity, facilitating more efficient allocations of analyst time (e.g., staffing decisions). Second, the VE recruitment classification methods developed in our prior study [35] could serve as a pre-screening step to identify recruitment messages, thus reducing the volume of documents requiring human attention. Used in tandem, these methods could reduce costs associated with manual analysis of online content. More generally, our classification and forecasting methods could be used within a VE cyber-recruitment identification and tracking methodology that would facilitate the study of recruitment efforts and the membership dynamics of violent organizations. Such a methodology might measure the effects of recruitment and counter-recruitment efforts on new membership by correlating specific recruitment activities and current events with changes in the VE population of a community. As a future research path, this proposed methodology requires (1) an automated system for classifying whether a forum user is a member of a violent extremist group, and (2) additional time series methods for analyzing the relationships between recruitment and membership over time. Identifying connections between VE recruitment levels and different events is another potentially useful extension of this research. Besides gaining a better understanding of the impact that different events have on recruitment dynamics, being able to connect large scale events, like the Syrian conflict, or more abrupt events, like the 2014 shootings in Canada's Parliament, to VE recruitment may also be useful for improving automated prediction techniques like the ones presented.

In the future, our forum post classification methods could be improved by including support for non-English languages (e.g., via automatic translation or direct use of non-English features). Support for other languages is an important task since many VE groups operate in non-English-speaking online communities. Our forecasting methods might be improved by generating additional latent topic predictors (we only used 30 in our experiments) and employing more sophisticated regression methods like gradient boosting or random forests. Future work on the forecasting task should also consider more advanced natural language processing techniques like supervised latent Dirichlet allocation (sLDA) [49], which jointly models the topic distributions and regression coefficients, and dynamic topic models [32], [33], which account for time-varying topics. The use of dynamic topic models would be feasible in combination with the ARIMA methods we have explored in the present research; however, our experience has been that even basic LDA is quite demanding computationally, and the addition of temporal variables to the graphical LDA structure would increase the running time of such an approach. Employing more sophisticated time series and regression techniques, including nonlinear models like ARCH or deep neural networks, could also improve upon our forecasting results. These models would be good candidates to replace our PCR-based approach, which showed significant evidence of nonlinearity per the BDS test. Additionally, a cointegration analysis would be useful for determining potential causation between the textual content of posts and VE recruitment activity. We note, though, that the predictions we have generated are useful, independent of the causal connection between predictors and responses (see above discussion on analyst staffing strategies).

Our work has focused on the domain of violent extremism; however, the combination of topic analysis and time series forecasting is applicable to other domains that generate digitized, timestamped text. For example, companies might be interested in forecasting future sentiment values within online reviews of their products. The sentiment values (positive and negative) are analogous to our recruitment/non-recruitment rating of forum posts, and the hypothesis in each domain is that topics of discussion at time $t - 1$ are predictive of response values at time t. The key contribution of the present work is to combine quantitative measurements of discussion topics (LDA) with time series forecasting methods (ARIMA and ETS) to improve upon the predictive accuracy of time-series-only approaches.

References

[1] L. A. Overbey, G. McKoy, J. Gordon, and S. McKitrick, "Automated sensing and social network analysis in virtual worlds," in *Proc. IEEE Int. Conf. Intell. Secur. Inform. (ISI)*, Vancouver, BC, Canada, May 2010, pp. 179–184.

[2] R. Torok, "'Make a bomb in your mums kitchen': Cyber recruiting and socialisation of 'White Moors' and home grown Jihadists," in *Proc. 1st Austral. Counter Terrorism Conf.*, Nov. 2010, pp. 54–61.

[3] W. V. Fitzgerald. (Jun. 2010). *Interview With Westboro Baptist Church: Hate Name God.* [Online]. Available: http://www.digitaljournal.com/article/293364

[3] http://ptl.sys.virginia.edu/msg8u/ve_recruitment_data.zip

[4] M. Rogers, “The psychology of cyber-terrorism,” in *Terrorists, Victims and Society: Psychological Perspectives on Terrorism and Its Consequences*. Chichester, U.K.: Wiley, 2003, pp. 77–92.

[5] S. O’Rourke, “Virtual radicalisation: Challenges for police,” in *Proc. 8th Austral. Inf. Warfare Secur. Conf.*, Dec. 2007, pp. 29–35.

[6] S. Mandal and E.-P. Lim, “Second life: Limits of creativity or cyber threat?” in *Proc. IEEE Conf. Technol. Homeland Secur.*, May 2008, pp. 498–503.

[7] R. R. Tomes, “Waging war on terror relearning counterinsurgency warfare,” *Parameters*, vol. 34, no. 1, pp. 16–28, 2004.

[8] F. Gutiérrez, “Recruitment in a civil war: A preliminary discussion of the Colombian case,” 2006.

[9] M. Humphreys and J. M. Weinstein, “Who fights? The determinants of participation in civil war,” *Amer. J. Political Sci.*, vol. 52, no. 2, pp. 436–455, 2008.

[10] M. I. Lichbach, *The Rebel’s Dilemma*. Ann Arbor, MI, USA: Univ. Michigan Press, 1998.

[11] K. Peters and P. Richards, “‘Why we fight’: Voices of youth combatants in Sierra Leone,” *J. Int. African Inst.*, vol. 68, no. 2, pp. 183–210, Apr. 1998.

[12] J. M. Weinstein, *Inside Rebellion: The Politics of Insurgent Violence*. New York, NY, USA: Cambridge Univ. Press, 2007.

[13] R. D. Petersen, *Resistance and Rebellion: Lessons From Eastern Europe*. New York, NY, USA: Cambridge Univ. Press, 2001.

[14] S. L. Popkin, “The rational peasant,” *Theory Soc.*, vol. 9, no. 3, pp. 411–471, 1980.

[15] J. C. Scott, *The Moral Economy of the Peasant: Rebellion and Subsistence in Southeast Asia*. London, U.K.: Yale Univ. Press, 1976.

[16] E. J. Wood, *Insurgent Collective Action and Civil War in El Salvador*. New York, NY, USA: Cambridge Univ. Press, 2003.

[17] R. W. McGehee, *Deadly Deceits: My 25 Years in the CIA*, Z. Sklar, Ed. New York, NY, USA: Sheridan Square Pub., 1983.

[18] M. Conway, “Terrorism and the Internet: New media—New threat?” *Parliamentary Affairs*, vol. 59, no. 2, pp. 283–298, 2006.

[19] R. Torok, “Developing an explanatory model for the process of online radicalisation and terrorism,” *Secur. Inform.*, vol. 2, no. 1, pp. 1–10, 2013.

[20] L. Bowman-Grieve, “A psychological perspective on virtual communities supporting terrorist & extremist ideologies as a tool for recruitment,” *Secur. Inform.*, vol. 2, no. 1, pp. 1–5, 2013.

[21] E. F. Kohlmann, “Al-Qaida’s MySpace: Terrorist recruitment on the Internet,” *CTC Sentinel*, vol. 1, no. 2, pp. 8–9, 2008.

[22] L. A. Overbey *et al.*, “Virtual DNA: Investigating cyber-behaviors in virtual worlds,” Space and Naval Warfare System Center Atlantic, Charleston, SC, USA, Tech. Rep. 33-09E, 2009.

[23] G. S. McNeal, “Cyber embargo: Countering the Internet jihad,” *Case Western Reserve Univ. J. Int. Law*, vol. 39, pp. 789–826, 2008.

[24] H. Chen, S. Thoms, and T. Fu, “Cyber extremism in Web 2.0: An exploratory study of international Jihadist groups,” in *Proc. IEEE Int. Conf. Intell. Secur. Inform. (ISI)*, Taipei, Taiwan, Jun. 2008, pp. 98–103.

[25] M. Yang, M. Kiang, H. Chen, and Y. Li, “Artificial immune system for illicit content identification in social media,” *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 2, pp. 256–269, 2012.

[26] A. Basu, “Social network analysis of terrorist organizations in India,” in *Proc. Conf. North Amer. Assoc. Comput. Soc. Org. Sci. (NAACSOS)*. Notre Dame, IN, USA, 2005, pp. 26–28.

[27] K. M. Carley, “Destabilization of covert networks,” *Comput. Math. Org. Theory*, vol. 12, no. 1, pp. 51–66, 2006.

[28] M. Chau and J. Xu, “Using Web mining and social network analysis to study the emergence of cyber communities in blogs,” in *Terrorism Informatics*. New York, NY, USA: Springer-Verlag, 2008, pp. 473–494.

[29] J. Diesner and K. M. Carley, “Using network text analysis to detect the organizational structure of covert networks,” in *Proc. Conf. North Amer. Assoc. Comput. Soc. Org. Sci. (NAACSOS)*, Pittsburgh, PA, USA, 2004, pp. 1–6.

[30] Z. Chen, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, “Identifying intention posts in discussion forums,” in *Proc. NAACL-HLT*, Jun. 2013, pp. 1041–1050.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[32] X. Wang and A. McCallum, “Topics over time: A non-Markov continuous-time model of topical trends,” in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 424–433. [Online]. Available: http://dl.acm.org/citation.cfm?id=1150450

[33] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120. [Online]. Available: http://dl.acm.org/citation.cfm?id=1143859

[34] T. Fu, A. Abbasi, and H. Chen, “A focused crawler for dark Web forums,” *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 6, pp. 1213–1231, 2010.

[35] J. R. Scanlon and M. S. Gerber, “Automatic detection of cyber-recruitment by violent extremists,” *Secur. Inform.*, vol. 3, no. 1, pp. 1–10, Aug. 2014.

[36] H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, and G. Weimann, “Uncovering the dark Web: A case study of Jihad on the Web,” *J. Amer. Soc. Inf. Sci. Technol.*, vol. 59, no. 8, pp. 1347–1359, 2008.

[37] Artificial Intelligence Laboratory, University Of Arizona. (2014). *Dark Web Forum Portal: Ansar Al Jihad Network English Website*. [Online]. Available: http://cri-portal.dyndns.org

[38] Google Inc. (2014). *Google Translate*. [Online]. Available: http://translate.google.com/

[39] M. S. Gerber, “Predicting crime using Twitter and kernel density estimation,” *Decision Support Syst.*, vol. 61, pp. 115–125, May 2014.

[40] I. Feinerer and K. Hornik. (2014). *tm: Text Mining Package, R Foundation for Statistical Computing, R Package Version 0.5-10*. [Online]. Available: http://CRAN.R-project.org/package=tm

[41] T. P. Jurka, L. Collingwood, A. E. Boydstun, E. Grossman, and W. van Atteveldt. (2014). *RTextTools: Automatic Text Classification Via Supervised Learning, R Package Version 1.4.2*. [Online]. Available: http://CRAN.R-project.org/package=RTextTools

[42] R Core Team. (2014). *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*. [Online]. Available: http://www.R-project.org/

[43] P. Whittle, *Hypothesis Testing in Time Series Analysis*, vol. 4. Uppsala, Sweden: Almqvist & Wiksells, 1951.

[44] R. J. Hyndman *et al.* (2014). *Forecast: Forecasting Functions for Time Series and Linear Models, R Package Version 5.3*. [Online]. Available: http://CRAN.R-project.org/package=forecast

[45] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, “A state space framework for automatic forecasting using exponential smoothing methods,” *Int. J. Forecast.*, vol. 18, no. 3, pp. 439–454, 2002.

[46] B.-H. Mevik, R. Wehrens, and K. H. Liland. (2013). *PLS: Partial Least Squares and Principal Component Regression, R Package Version 2.4-3*. [Online]. Available: http://CRAN.R-project.org/package=pls

[47] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006.

[48] W. H. Webster, D. E. Winter, A. L. Steel, Jr., W. M. Baker, R. J. Bruemmer, and K. L. Wainstein, “Final report of the William H. Webster Commission on the Federal Bureau of Investigation, counterterrorism intelligence, and the events at Fort Hood, Texas on November 5, 2009,” FBI Headquarters, Washington, DC, USA, Tech. Rep., 2012.

[49] D. M. Blei and J. D. McAuliffe, “Supervised topic models,” in *Proc. NIPS*, vol. 7. 2007, pp. 121–128.

Jacob R. Scanlon received the B.S. degree in systems and information engineering and the M.S. degree from the University of Virginia, in 2014. He was an Operations Research Analyst with Booz Allen Hamilton. He is currently a Data Scientist with the data warehousing and analytics firm Teradata.

Matthew S. Gerber received the Ph.D. degree in computer science from Michigan State University, in 2011. He is currently an Assistant Professor with the Department of Systems and Information Engineering, University of Virginia at Charlottesville. He has worked in the areas of natural language processing, crime prediction, and mobile sensing.